Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230921

Studies advanced in chatbots based on deep learning

Lixin Li

Franklin and Marshall College, Lancaster, Pennsylvania, 17603, US

lli1@fandm.edu

Abstract. Chatbots have always been a hot research topic in the field of human-computer interaction research, which aims to build a conversational intelligent response model to simulate human dialogue. Thanks to the rapid development of natural language processing technology and the continuous accumulation of dialogue data, the research of chat robots have made remarkable progress, which has gradually been widely used in various fields such as e-commerce and smart home. According to different technical frameworks, existing chatbots are mainly divided into two types: retrieval chatbots and generative chatbots. As the primary means of implementing chatbots in the industry, retrieval chatbots have smooth responses and low computational resource consumption. In contrast, generative chatbots do not require a predefined knowledge base and can dynamically generate responses based on the dialogue content. In this paper, focusing on the above two types of frameworks, we introduce the latest research progress in the field of deep learning-based chatbots in detail, including the representative algorithms and corresponding pipelines. Second, we compare the performance of representative algorithms on different datasets. We also summarize the problems chatbot technology research faces and give an outlook on its future development trends.

Keywords: Natural Language Processing, Chatbot, Deep Learning.

1. Introduction

Chatbots have always been a hot research topic in the field of human-computer interaction, which aims to respond naturally to human input content (mainly text or voice), for example, to complete a smooth and complex logical dialogue [1]. With the rapid development of artificial intelligence technology, chatbots have received extensive attention from industry and academia in recent years. They have been successfully applied in different fields, such as business, education, and information. The emergence of chatbots can save a lot of labor costs, such as artificial intelligence customer service. They categorize user queries and then provide appropriate responses. For example, suppose a buyer asks whether a merchant can return the product for free. In that case, the robot customer service will reply with a predefined answer after evaluating the inquiry, such as they can return the product for free, saving a lot of time and labor costs.

Chatbot research has a long history, traced back to Eliza, which was developed by researchers in the MIT laboratory in the 1960s [2]. As a psychotherapist, Eliza relies on templates and returns user comments in the form of questions. If what the user says matches the templates written, you can get a good reply; otherwise, you will get some universal replies. At the end of the 20th century, rule-based

DOI: 10.54254/2755-2721/6/20230921

chatbots became a research hotspot, and the representative ones are Parry and Alicebot. However, human languages are varied, and relying solely on template technology cannot enumerate all situations, so the development of chatbots was once at a bottleneck.

After entering the 21st century, with the development of machine learning technology and more available Internet dialogue materials, data-driven chat robot technology has become more mature. According to the difference in design thinking, the most representative chat robots can be divided into retrieval-based chatbots and generation-based chatbots. Retrieval-based chatbots refer to the use of information retrieval technology to match a pre-stored dialogue material as a reply to a user's conversation request. The retrieval model is relatively simple, and its effect mainly depends on the knowledge base extraction, retrieval technology, and sorting features. Generation-based chatbots refer to the use of natural language generation technology to automatically reply to user conversation requests. The generative model relies on a large amount of training data to learn a very powerful representation ability of the semantic features of natural language, which can dynamically generate natural and logical responses according to the input content. Existing mainstream generation-based chatbots borrow from sequence-to-sequence models that have been successful when it comes to machine translation.

Both generative and retrieval-based chatbots have advantages and disadvantages. However, thanks to the smaller computing overhead, smooth response, and a large amount of information, search-based chatbots have become the mainstream solution in the industry. In this paper, we introduce the latest research programs of chatbots through a large amount of literature research and analysis around the above two types of representative frameworks. Specifically, Chapter 2 introduces the respective characteristics and representative algorithms of the two types of frameworks. Sheet 3 presents relevant public datasets and compares the performance of different algorithms. Finally, we explore the remaining research questions on this topic and look forward to the future development trend of retrieval chatbots.

2. Method

2.1. Retrieval-based chatbot models

Retrieval-based chatbot models use a predefined knowledge base to formulate a set of structured policies and model rules. Through text matching and ranking learning technology, retrieval-based chatbots search the discourse corpus for the best response to the current input. Candidate index retrieval, similarity feature computation, and ranking learning are the three main elements of a retrieval chatbot. The candidate index retrieval module pre-assembles massive amounts of human dialogue data and arranges it in a question-answer manner [3]. Then the indexing technologies in information retrieval are used to index these conversations for fast retrieval by online modules. The similarity feature computation module trains a variety of various reply selection models beforehand using a significant volume of unlabelled or sparsely labeled data. Using the characteristics provided by the similarity feature computation module and the labels of the labeled data, the ranking learning module develops the ranking learning model.

In the inference stage, for given user input, a retrieval chatbot first quickly retrieves several candidates' replies from large-scale dialogue data. The candidate index retrieval module employs the inverted list or vector retrieval approach to swiftly respond and recall since the dialogue data is frequently in the tens of millions. These response selection models then give each potential response a score and combine the scores into a vector. To achieve the final ranking of candidate replies, the ranking learning module incorporates these features. The two modules of ranking learning and similarity feature computation have very strict accuracy requirements. Some more complicated models are frequently utilized in the hopes that the most pertinent candidate replies can be prioritized at the top of the ranking results. It is crucial to remember that the module for the similarity feature computation must take the current situation into account, as well as the degree to which the candidate replies are similar on several dimensions, including relevance, logical consistency, and stylistic consistency.

Ji et al. designed a retrieval-based chatbot using a large amount of short dialogue data on Weibo, which uses the traditional TF-IDF technology to obtain potential solutions from the knowledge base and

DOI: 10.54254/2755-2721/6/20230921

then extract multiple features from the candidate answers to train the targeted ranking model to sort the candidate answers [4]. Yan et al. used unstructured documents as the knowledge base, adopted BM25 retrieval technology, and extracted different features to train the ranking model [5]. Chatbots of this type of retrieval model cannot reconstruct the answer results, and knowledge-based retrieval relies on empirical knowledge and does not generate new text. For the problems that have been encountered, the result is relatively good, but for the issues that have yet to be encountered, it cannot give good results. All the answers it provides already exist, and it is more dependent on retrieval technology and knowledge library.

2.2. Generation-based chatbot models

The chatbots based on the generation model do not require a predefined knowledge base, which aims to build natural user interfaces that depending on user intent, understand context and meaning from unstructured natural language user input [6]. Generation-based chatbots understand what the user says and instructs the chatbot to answer appropriately based on the purpose. The advantages of this form of chatbot are precisely what the disadvantages of the previous type of chatbot are. They are more human-like and may respond better to orders given to them. The problem is that training such bots frequently necessitates considerable data collection and a lengthy training period.

Most Generation-based chatbots usually follow the framework of machine translation. Ritter et al. used a large amount of Weibo chat data and statistical machine translation technology to translate questions (equivalent to the source language of translation) into answers (equivalent to the target language of translation) [7]. Vinyals et al. used neural network machine translation as a chat robot [8]. Shang et al. also used a large amount of Weibo chat data, which adopted a Sequence-to-Sequence framework RNN network and constructed a Neural Responding Machine (NRM) [9]. Serban et al. employed a layered network to generate chatbots [10]. Chatbots of this type of model can create answers to arbitrary questions. Still, their responses are often unreasonable and unnatural, and the training of their models requires a large amount of question-and-answer data.

Each of the two ways outlined above has benefits and drawbacks. Because they merely choose an answer from a list of predetermined responses, retrieval-based techniques do not produce syntax mistakes. They cannot, however, handle unexpected circumstances since no relevant predetermined solution exists in the pool of predefined replies. For the same reason, these models are incapable of comprehending earlier settings. These models cannot reference information such as the names of locations, individuals, or anything else discussed previously in the dialogue. On the other hand, generative models may recall past knowledge and are thus more "intelligent". This improves the human-computer interface. These models, however, are challenging to train.

3. Performance comparison of representative methods

3.1. Classic datasets

In this section, we first introduce standard public datasets for chatbot training, which mainly include: Ubuntu Dialogue Corpus, Santa Barbara Corpus of Spoken American English, and The SMS Corpus of the National University of Singapore.

The Ubuntu Dialogue Corpus dataset comprises approximately a million two-person talks taken from Ubuntu chats and used to obtain technical help for various Ubuntu-related topics. Each conversation lasts eight rounds on average, with a minimum of 3 rounds. All talks are held in written (rather than audio) form. The total dataset includes 930,000 talks and more than 100,000,000 words.

Barbara Corpus of Spoken American English dataset comprises roughly 249,000 words, as well as transcriptions, audio, and timestamps that correlate transcriptions and audio at the level of individual intonation units.

The Department of Computer Science at the National University of Singapore collected SMS (Short Message Service) conversations for study and created the SMS Corpus of the National University of Singapore. On March 9, 2015, 67093 SMS messages were collected from the corpus. Most of the

DOI: 10.54254/2755-2721/6/20230921

Singaporeans who are sending these mails are university students. Volunteers who knew their contributions would be made public provided the messages.

3.2. Evaluation metrics

Several standard metrics are adopted to evaluate the performance of different chatbots. Matthews invented the MCC (Matthews correlation coefficient) in 1975 to compare chemical structures. It was later reintroduced in 2020 as a standard machine learning performance indicator and naturally expanded to multi-class scenarios. However, due to inconsistent categorization findings, MCC cannot define or illustrate substantial swings. Many studies consider the ratio of properly categorized samples to the total number of pieces the most realistic performance statistic. This statistic is known as "accuracy," and it, by definition, pertains to instances with more than two labels (the multi-class case). Accuracy is no longer taken into consideration when the data set is unbalanced (the number of samples in one class is much higher than the number of samples in the other classes). The F1 score is the most well-known member of the parametric family of F-measures, which is the harmonic mean of accuracy. The MCC provides more accurate and helpful findings when rather than accuracy and F1 ratings, binary classifications are evaluated. When assessing binary classification tasks, accuracy and F1 scores should come after the Matthews correlation coefficient.

3.3. Comparison of different algorithms

The comparison in this section focuses on the accuracy between models or which models are better by the values of PPL, Average BLEU, MAP, and MRR. In the following, we present two types of comparisons. The first one is the MAP and MRR used by Xiang et al. [11] and the other one is the PPL and BLEU used by Rashkin et al. [12] the table below collates the accuracy of the models used in chatbots.

MAP is "Mean Average Precision," a concept introduced by Baeza et al. in 1999 [13], and MRR is "Mean Reciprocal Rank," a concept mentioned by Voorhees et al. in 1999 [14].

The area under the P-R curve is referred to as MAP. The P-R curve, meanwhile, may be read as illustrating the connection between accuracy and recall at a given threshold value. Finally, MAP is the AP average. Perplexity (PPL) measures the degree of fit of a probability distribution or probability model to the sample, and the lower the perplexity, the more accurate the fit. Bi-Lingual Evaluation Understudy (BLEU) score assesses how closely the machine-translated text resembles a collection of professional reference translations and goes from 0 to 1.

Table 1. Comparison of different types of models.

Models	PPL	Avg BLEU	MAP(%)	MRR
Fine-Tuned [12]	21.24	6.27	-	-
Pretrained [12]	27.96	5.01	-	-
MULTITASK[12]	24.07	4.36	-	-
ENSEM-DM [12]	19.05	6.83	-	-
DAM [11]	-	-	0.550	0.601
MV-LSTM [11]	-	-	0.498	0.538
DL2R [11]	-	-	0.488	0.527
SMNdynamic [11]	-	-	0.529	0.569

As we can see from this table, the model DAM has the highest MAP and MRR. MAP reaches a value of 55.0%, which is the highest among all models. The MRR of DAM is also the highest among all models. So we can say that DAM performs better than MV-MSTM, DL2R, and SMN.

DOI: 10.54254/2755-2721/6/20230921

At the same time, we can see that the PPL of the ENSEM-DM model is the lowest among all models, which is 19.05. This means that the ENSEM-DM model has the highest model fit and is the most accurate. From these two sets of data, it can be seen that PPL, average BLEU, MAP, and MRR can accurately distinguish different models in terms of model accuracy. They are only different in variance from the data, but they are all very effective in judging the models.

4. Discussion

Although the current chatbot has achieved preliminary results, it still faces some serious problems and severe challenges.

Multi-modal dialogue. Existing chatbots only consider textual information, but conversations between people often contain more modal information, and this information is critical to the understanding of the conversation. For example: "How can I be as smart as you?" If the tone is sarcastic, the speaker is holding a negative emotion; while if the tone is cheerful, the speaker is holding a positive emotion. In addition to the tone of voice, facial expressions and body language can also better aid in the understanding of multiple rounds of conversation. Therefore, when building a chatbot, introducing multi-modal dialogue can better help the understanding of multiple rounds of dialogue, thereby improving the user experience of the chatbot.

Long sentence. If the user is inputting extensive text, the bot may have too much input and be unable to focus, resulting in misclassification or classification failure. This may cause our chatbot to respond incorrectly. To address this issue, we need a large enough dataset for the chatbot to be trained with lengthy text, which increases the likelihood of successful categorization.

Model reasoning. Retrieval chatbots mainly consider the context and relevance of candidate responses, which do not perform well when they need to make simple model inferences. For example, for "here is six hours later than New York. It is five o'clock in the afternoon in New York", it is often difficult for the machine to deduce the current time. In addition, the machine cannot reason well about the user's emotional polarity towards some things, so it cannot start a dialogue on the user's attitude and emotions. Therefore, robots can still only make simple, relevant replies and cannot reason about conversations like humans.

5. Conclusion

In this paper, focusing on two types of frameworks: retrieval-based chatbots and generative-based chatbots, we introduce the latest research progress in the field of chatbots based on deep learning, including representative algorithms and corresponding pipelines. Second, we compare the performance of representative algorithms on different datasets. We also summarize the problems facing chatbot technology research and look forward to its future development trends.

References

[1] Adamopoulou, Eleni, and Lefteris Moussiades. "An Overview of Chatbot Technology." *IFIP Advances in Information and Communication Technology*, 2020, pp. 373–383., https://doi.org/10.1007/978-3-030-49186-4_31.

[2] Weizenbaum, Joseph. "Eliza—a Computer Program for the Study of Natural Language Communication between Man and Machine (1966)." Ideas That Created the Future, 2021, pp. 271–278., https://doi.org/10.7551/mitpress/12274.003.0029.

[3] Akma, Nahdatul, et al. "Review of Chatbots Design Techniques." International Journal of Computer Applications, vol. 181, no. 8, 2018, pp. 7–10., https://doi.org/10.5120/ijca2018917606.

[4] Ji Z, Lu Z, Li H. "An information retrieval approach to short text conversation[J]". arXiv preprint arXiv:1408.6988, 2014.

[5] Yan Z, Duan N, Bao J, et al. Docchat. "An information retrieval approach for chatbot engines using unstructured documents[C]"//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 516-525.

DOI: 10.54254/2755-2721/6/20230921

[6] Jung, Sangkeun. “Semantic Vector Learning for Natural Language Understanding.” Computer Speech & Language, vol. 56, 2019, pp. 130–145., https://doi.org/10.1016/j.csl.2018.12.008.
[7] Ritter A, Cherry C, Dolan B. “Data-driven response generation in social media[C]”//Empirical Methods in Natural Language Processing (EMNLP). 2011.
[8] Vinyals O, Le Q. “A neural conversational model[J]”. arXiv preprint arXiv:1506.05869, 2015.
[9] Shang L, Lu Z, Li H. “Neural responding machine for short-text conversation[J]”. arXiv preprint arXiv:1503.02364, 2015.
[10] Serban I, Sordoni A, Bengio Y, et al. “Building end-to-end dialogue systems using generative hierarchical neural network models[C]”//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1).
[11] Zhou, Xiangyang, et al. “Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network.” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, https://doi.org/10.18653/v1/p18-1103.
[12] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open- domain conversation models: A new benchmark and dataset. In Proceedings of the 57th Confer- ence of the Association for Computational Linguis- tics, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
[13] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. Modern information retrieval, volume 463. ACM press New York.
[14] Ellen M Voorhees et al. 1999. The trec-8 question an- swering track report. In Trec, pages 77–82.